Rethinking the Reference model in RLHF

YiMing Liu, letusgo126@126.com

March 2025

1 贡献

- 在 appendix A 提出了分类损失中 KL 散度梯度的无偏估计方法，揭示了 k1 与 k2 估计量之间的等价关系，并在 section 2.2 从理论上证明了 k2 作为 KL 损失函数的优越性，其梯度估计相较于 k3 方法具有更高的无偏性。
- 在 section 2.1.4, 推导了 PPO 算法的原始奖励公式与优化目标，为策略优化提供了清晰的理论框架，澄清了现有文献中若干模糊表述。
- 在 section 3.2 中，提出了一种新颖的 KL 惩罚项集成方法，成功应用于 Diffusion RLHF 框架下的 DDPO 算法，解决了传统方法依赖外部数据集引入扩散预训练损失的局限性，显著降低了分布偏移风险。
- 在 section 3.3 首次指出了当前 DiffusionDPO 理论中存在的关键错误

2 LLM RLHF

2.1 基于策略梯度方法的 PPO 奖励函数推导

2.1.1 RLHF 目标函数建模

基于人类反馈的强化学习（RLHF）的核心目标可建模为以下优化问题：

$$\arg\max_{\theta} J = \underbrace{\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_\theta(y|x)}\left[r(x,y)\right]}_{\text{奖励最大化项}} - \beta\underbrace{\left[\mathbb{D}_{\text{KL}}\left(\pi_\theta(y|x)\,||\,\pi_{\text{ref}}(y|x)\right)\right]}_{\text{策略约束项}} \quad (1)$$

其中 β 为温度系数，控制策略偏离参考模型 π_{ref} 的程度。该目标函数包含两个关键部分：奖励期望的最大化和 KL 散度约束的平衡。

2.1.2 目标函数分解与分析

首先考虑奖励相关项的损失函数构造。根据策略梯度定理，可建立奖励项的损失函数为：

$$-\mathcal{L}_R(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_{\theta_{old}}(y|x)}\left[r(x,y)\log\pi_\theta(y|x)\right] \tag{2}$$

其中 $\pi_{\theta_{old}}$ 为采样时的模型参数，π_θ 为 RL 训练时的模型参数，如果采样一次训练一次，二者在数值上相等。

结合 KL 散度约束项 $\mathcal{L}_{\mathrm{KL}}(\pi_\theta, \pi_{ref})$，完整损失函数可表示为：

$$L_{total} = -\left(\mathcal{L}_R(\theta) - \beta\mathcal{L}_{\mathrm{KL}}(\pi_\theta, \pi_{ref})\right) \tag{3}$$

2.1.3 梯度推导过程

对目标函数进行梯度分析时，需要分别处理两个组成部分：

1. **奖励项梯度**：

$$\nabla_\theta\mathcal{L}_R(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_{\theta_{old}}(y|x)}\left[r(x,y)\nabla_\theta\log\pi_\theta(y|x)\right] \tag{4}$$

2. **KL 散度项梯度**（具体推导见引理 eq. (48)）：

$$-\nabla_\theta\mathcal{L}_{\mathrm{KL}_t} = -\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_{\theta_{old}}(y|x)}\left[\log\frac{\pi_{\theta_{old}}(y|x)}{\pi_{ref}(y|x)}\cdot\nabla_\theta\log\pi_\theta(y|x)\right] \tag{5}$$

将二者结合可得总梯度：

$$\begin{aligned}\nabla_\theta J &= \nabla_\theta\mathcal{L}_R - \beta\nabla_\theta\mathcal{L}_{\mathrm{KL}_t} \\ &= \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_{\theta_{old}}(y|x)}\left[\left(r(x,y) - \beta\log\frac{\pi_{\theta_{old}}(y|x)}{\pi_{ref}(y|x)}\right)\nabla_\theta\log\pi_\theta(y|x)\right]\end{aligned} \tag{6}$$

2.1.4 等效奖励函数构造与 PPO 算法的 reward 函数

由公式 eq. (6) 可以重新定义新的优化目标，也就是 PPO 的优化目标：

$$\text{原问题}: \arg\max_\theta J' = \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_\theta(y|x)}\left[r(x,y) - \beta\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}\right] \tag{7}$$

该形式将 KL 正则项内化为奖励函数的组成部分，从而实现了优化目标的重新参数化。

容易得出 PPO 算法的瞬时奖励函数为：

$$\tilde{r}(x,y) := r(x,y) - \beta\log\frac{\pi_{\theta_{old}}(y|x)}{\pi_{\mathrm{ref}}(y|x)} \tag{8}$$

2.2 KL 惩罚项的数学性质分析

2.2.1 奖励函数中 KL 惩罚项的适用性条件

在 PPO 算法设计中，仅特定形式的 KL 散度估计量适合作为奖励函数的惩罚项。根据理论分析，k1 估计量具有数学适用性，其表达式如 eq. (8) 所示。然而，k2 和 k3 估计量由于内在的数学性质缺陷，不适合作为惩罚项，原因如下：

考虑 k2 和 k3 的估计值 kl 具有恒正特性，其梯度方向始终满足：

$$-\nabla_\theta \mathrm{KL}(\pi_\theta \| \pi_{ref}) = -kl \nabla_\theta \log \pi_\theta(y|x) \tag{9}$$

此时无论 $\pi_\theta(y|x)$ 与 $\pi_{ref}(y|x)$ 的概率分布关系如何，梯度更新方向都会强制降低当前策略 π_θ 生成任意动作 y 的概率。这种单向的惩罚机制本质上违背了策略优化算法的基本设计原则，即应建立方向可调的奖惩机制来引导策略改进。

2.2.2 KL 惩罚项的损失函数适配性

k2 估计量的有效性 如 eq. (49) 和 eq. (49) 所示，k2 估计量能够构造有效的 KL 惩罚损失函数。其数学本质是通过 KL 散度的对称性设计，建立双向调节机制：当 π_θ 偏离 π_{ref} 时，梯度方向会根据偏离方向自动调整，既防止策略过度偏离，又保留必要的优化自由度。

k1 估计量的失效机理 k1 对应的损失函数形式为：

$$\mathcal{L}_{\mathrm{KL}_t}(\pi_\theta, \pi_{ref}) = \log \pi_\theta(y|x) - \log \pi_{ref}(y|x) \tag{10}$$

其梯度表达式揭示本质缺陷：

$$-\nabla_\theta \mathcal{L}_{\mathrm{KL}_t} = -\nabla_\theta \log \pi_\theta(y|x) \tag{11}$$

该梯度恒指向降低当前策略概率的方向，形成类似极大似然估计的反向约束。这种单向作用机制会导致策略网络的概率输出持续衰减，最终引发模型坍缩问题。

k3 估计量的近似特性 k3 构造的损失函数具有特殊形式：

$$\mathcal{L}_{\mathrm{KL}_t} = \frac{\pi_{ref}(y|x)}{\pi_\theta(y|x)} - \log \frac{\pi_{ref}(y|x)}{\pi_\theta(y|x)} - 1 \tag{12}$$

对应的梯度表达式揭示其近似本质：

$$-\nabla_\theta \mathcal{L}_{\mathrm{KL}} = \left(\frac{\pi_{\mathrm{ref}}(y|x)}{\pi_{\theta_{old}}(y|x)} - 1 \right) \nabla_\theta \log \pi_\theta(y|x) \tag{13}$$

令 $x = \pi_{\mathrm{ref}}/\pi_\theta$，k2 与 k3 的梯度可对比表示为：

$$\text{k2 梯度} : \log x \cdot \nabla_\theta \log \pi_\theta \tag{14}$$

$$\text{k3 梯度} : (x-1) \cdot \nabla_\theta \log \pi_\theta \tag{15}$$

在策略邻域 ($x \approx 1$) 进行泰勒展开时，$\log x \approx x - 1$，此时 k3 梯度构成 k2 梯度的线性近似。但这种近似具有两个关键缺陷：1. **有偏性**：当策略显著偏离参考策略 (x 远离 1, 也就是训练后期 π_{ref} 离 π_θ 较远) 时，近似误差呈非线性增长 2. **非对称性**：$x - 1$ 对 $\pi_\theta > \pi_{ref}$ 和 $\pi_\theta < \pi_{ref}$ 的响应特性不对称。

3 Diffusion RLHF

3.1 Diffusion RLHF 之 DDPO

π 表示概率, 常用符号为 p。

3.1.1 扩散模型基础

Denoising Diffusion Probabilistic Models (DDPM) 的逆向过程采样公式定义为：

$$\pi_\theta \left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{c} \right) = \mathcal{N} \left(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta \left(\boldsymbol{x}_t, \boldsymbol{c}, t \right), \sigma_t^2 \mathbf{I} \right) \tag{16}$$

其具体采样形式为：

$$\boldsymbol{x}_{t-1} = \boldsymbol{\mu}_\theta \left(\boldsymbol{x}_t, \boldsymbol{c}, t \right) + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N} \left(\mathbf{0}, \mathbf{I} \right) \tag{17}$$

其中 $\boldsymbol{c}$ 为条件信息，σ_t^2 为预定义的方差参数。

3.1.2 概率密度与梯度推导

当方差固定时，高维高斯分布的概率密度函数可展开为：

$$\pi_{\boldsymbol{\theta}} \left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{c} \right) = \frac{1}{(2\pi\sigma_t^2)^{d/2}} \exp \left(-\frac{\| \boldsymbol{x}_{t-1} - \boldsymbol{\mu}_\theta \left(\boldsymbol{x}_t, \boldsymbol{c}, t \right) \|^2}{2\sigma_t^2} \right) \tag{18}$$

其中 d 为潜变量维度。取对数概率得：

$$\log \pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{c}\right) = -\frac{d}{2}\ln(2\pi) - \frac{d}{2}\ln \sigma_t^2 - \frac{\|\boldsymbol{x}_{t-1} - \boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t, \boldsymbol{c}, t\right)\|^2}{2\sigma_t^2} \tag{19}$$

对模型参数 θ 求梯度时，前两项为常数项，故梯度表达式简化为：

$$\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{c}\right) = -\frac{1}{2\sigma_t^2}\nabla_{\boldsymbol{\theta}}\|\boldsymbol{x}_{t-1} - \boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t, \boldsymbol{c}, t\right)\|^2 \tag{20}$$

3.1.3 Diffusion 强化学习目标函数

DDPO（Diffusion Policy Optimization）的优化目标定义为:

$$\arg\max_\theta J = \underbrace{\mathbb{E}_{c\sim\mathcal{D}, x_0\sim\pi_\theta(\cdot|\cdot,c)}\left[r(x_0, \boldsymbol{c})\right]}_{\text{奖励最大化项}} \tag{21}$$

DDPO 的梯度函数定义为：

$$\nabla_{\boldsymbol{\theta}} J = \nabla_{\boldsymbol{\theta}}\mathcal{L}_{DDPO} = \mathbb{E}_{\boldsymbol{c}\sim D, t\sim U(0,T)}\left[r\left(\boldsymbol{x}_0, \boldsymbol{c}\right)\nabla_{\boldsymbol{\theta}}\log \pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{c}\right)\right] \tag{22}$$

展开后得到完整表达式：

$$\begin{aligned} -\mathcal{L}_{DDPO} &= \mathbb{E}_{\boldsymbol{c}\sim D, t\sim U(0,T)}\left[r\left(\boldsymbol{x}_0, \boldsymbol{c}\right)\log \pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{c}\right)\right] \\ &= \mathbb{E}_{\boldsymbol{c}\sim D, t\sim U(0,T)}\left[r\left(\boldsymbol{x}_0, \boldsymbol{c}\right)\left(\underbrace{-\frac{d}{2}\ln(2\pi) - \frac{d}{2}\ln\sigma_t^2}_{\text{常数项}} - \frac{\|\boldsymbol{x}_{t-1} - \boldsymbol{\mu}_\theta\|^2}{2\sigma_t^2}\right)\right] \end{aligned} \tag{23}$$

注意到常数项对优化无贡献，可得等价目标函数：

$$-\mathcal{L}_{DDPO_{equiv}} = \mathbb{E}_{\boldsymbol{c}\sim D, t\sim U(0,T)}\left[-r\left(\boldsymbol{x}_0, \boldsymbol{c}\right)\frac{\|\boldsymbol{x}_{t-1} - \boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t, \boldsymbol{c}, t\right)\|^2}{2\sigma_t^2}\right] \tag{24}$$

3.1.4 与预训练目标的关联

对比扩散模型预训练目标：

$$-\mathcal{L}_{pretrain}(\theta) = \mathbb{E}_{(\mathbf{x}_0,\mathbf{c})\sim p(\mathbf{x}_0,\mathbf{c}), t\sim\mathcal{U}\{0,T\}, \mathbf{x}_t\sim q(\mathbf{x}_t|\mathbf{x}_0)}\left[-\frac{\|\tilde{\boldsymbol{\mu}}(\mathbf{x}_0, t) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, \mathbf{c}, t)\|^2}{2\sigma_t^2}\right] \tag{25}$$

可知 DDPO 在预训练目标基础上引入了奖励加权机制，实现对齐优化。

3.1.5 重要性采样扩展

引入重要性采样比后，DDPO 改进形式为：

$$
\begin{aligned}
-\mathcal{L}_{DDPO_{IS}} &= \mathbb{E}_{\boldsymbol{c}\sim D, t\sim U(0,T)}\left[r\left(\boldsymbol{x}_0,\boldsymbol{c}\right)\frac{p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{c}\right)}{p_{\boldsymbol{\theta_{old}}}\left(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{c}\right)}\right] \\
&= \mathbb{E}_{\boldsymbol{c}\sim D, t\sim U(0,T)}\left[r\left(\boldsymbol{x}_0,\boldsymbol{c}\right)\exp\left(\frac{\|\mathbf{x_{t-1}}-\boldsymbol{\mu}_{\theta_{old}}\left(\boldsymbol{x}_t,\boldsymbol{c},t\right)\|^2}{2\sigma_t^2}-\frac{\|\mathbf{x_{t-1}}-\boldsymbol{\mu}_{\theta}\left(\boldsymbol{x}_t,\boldsymbol{c},t\right)\|^2}{2\sigma_t^2}\right)\right]
\end{aligned}
\tag{26}
$$

其中 θ_{old} 为旧策略参数，通过重要性采样比 $\frac{p_\theta}{p_{\theta_{old}}}$ 实现策略的稳定更新。可进一步采用 PPO-2 clip 策略稳定训练。

实现细节说明 注：在理论推导中，L2 范数平方对应的损失函数采用总和归约（reduction='sum'），而实际代码实现中通常采用均值归约（reduction='mean'）以增强数值稳定性。这种理论推导与工程实践的差异性在 DDPM 的标准实现中同样存在，其主要差异体现在梯度幅值的缩放比例，而不会改变优化方向。具体表现为：

- 理论推导：$\|\cdot\|^2$ 对应元素级平方和（sum）
- 工程实现：$\frac{1}{d}\|\cdot\|^2$ 对应元素级平方均值（mean）

其中 d 为特征维度，这种缩放操作通过保持梯度量级在合理范围内，有效避免了训练过程中的数值溢出问题。以上推导中的 μ 替换成 DDPM 中的 ϵ、Rectified Flow 中的 v 和 score sde 中的 score 都成立，推导省略。

3.2 基于参考模型的扩散强化学习对齐方法 (DDPO with Reference Model)

$$
\arg\max_{\theta} J = \underbrace{\mathbb{E}_{c\sim\mathcal{D},x_0\sim\pi_\theta(\cdot|c)}\left[r(x_0,c)\right]}_{\text{奖励最大化项}} - \beta\underbrace{\mathbb{E}_{c\sim\mathcal{D},t\sim U(0,T)}\left[\mathbb{D}_{\mathrm{KL}}\left(\pi_\theta(\cdot|x_t,c)\parallel\pi_{\mathrm{ref}}(\cdot|x_t,c)\right)\right]}_{\text{策略正则化项}}
\tag{27}
$$

根据 appendix B 的推导结果，KL 散度的期望可以显式表达为：

$$
\mathcal{L}_{DDPO_{KL}} = \mathbb{E}_{c\sim\mathcal{D},t\sim U(0,T)}\left[\frac{\|\boldsymbol{\mu}_\theta(x_t,c,t)-\boldsymbol{\mu}_{\mathrm{ref}}(x_t,c,t)\|^2}{2\sigma_t^2}\right] \tag{28}
$$

$$-\nabla \boldsymbol{\mu}_\theta(x_t,c,t)\mathcal{L}_{DDPO_{KL}} = \mathbb{E}_{c\sim\mathcal{D},t\sim U(0,T)}\left[\frac{\left(\boldsymbol{\mu}_{\text{ref}}(x_t,c,t)-\boldsymbol{\mu}_\theta(x_t,c,t)\right)}{\sigma_t^2}\right] \tag{29}$$

因此，完整的 DDPO 目标函数可分解为以下形式：

$$\begin{aligned}
\mathcal{L}_{DDPO_{total}} &= -\left(\mathcal{L}_{DDPO} - \beta\mathcal{L}_{DDPO_{KL}}\right) \\
&= -\mathbb{E}_{c\sim D,t\sim U(0,T)}\Big[r(x_0,c)\log\pi_\theta(x_{t-1}|x_t,c)\Big] \\
&\quad + \beta\mathbb{E}_{c\sim\mathcal{D},t\sim U(0,T)}\left[\frac{\|\boldsymbol{\mu}_\theta(x_t,c,t)-\boldsymbol{\mu}_{\text{ref}}(x_t,c,t)\|^2}{2\sigma_t^2}\right] \\
&= \mathbb{E}_{c\sim D,t\sim U(0,T)}\left[\frac{r(x_0,c)}{2\sigma_t^2}\|x_{t-1}-\boldsymbol{\mu}_\theta(x_t,c,t)\|^2\right. \\
&\quad \left. + \frac{\beta}{2\sigma_t^2}\|\boldsymbol{\mu}_\theta(x_t,c)-\boldsymbol{\mu}_{\text{ref}}(x_t,c,t)\|^2\right]
\end{aligned} \tag{30}$$

3.3 Diffusion RLHF 中 DiffusionDPO 推导的谬误分析

若直接沿用 PPO 的奖励函数形式（如 eq. (8) 所示），将推导出错误的损失函数。

$$\begin{aligned}
\mathcal{L}_{DDPO_{wrong}} &= -\mathbb{E}_{\boldsymbol{c}\sim D,t\sim U(0,T)}\left[\left(r\left(\boldsymbol{x}_0,\boldsymbol{c}\right)-\beta\log\frac{\pi_{\boldsymbol{\theta}_{old}}\left(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{c}\right)}{\pi_{\boldsymbol{ref}}\left(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{c}\right)}\right)\right. \\
&\qquad \left.\times\log\pi_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{c}\right)\right] \\
&= \mathbb{E}_{\boldsymbol{c}\sim D,t\sim U(0,T)}\left[\left(r\left(\boldsymbol{x}_0,\boldsymbol{c}\right)-\beta\left(\frac{\|\boldsymbol{x}_{t-1}-\boldsymbol{\mu}_{ref}\left(\boldsymbol{x}_t,\boldsymbol{c},t\right)\|^2}{2\sigma_t^2}-\frac{\|\boldsymbol{x}_{t-1}-\boldsymbol{\mu}_{\theta_{old}}\left(\boldsymbol{x}_t,\boldsymbol{c},t\right)\|^2}{2\sigma_t^2}\right)\right)\right. \\
&\qquad \left.\cdot\frac{\|\boldsymbol{x}_{t-1}-\boldsymbol{\mu}_\theta\left(\boldsymbol{x}_t,\boldsymbol{c},t\right)\|^2}{2\sigma_t^2}\right]
\end{aligned} \tag{31}$$

显然 $\mathcal{L}_{DDPO_{wrong}} \neq \mathcal{L}_{DDPO_{total}}$。而其根本原因在于 Diffusion 的 KL 惩罚梯度 eq. (76) 和 LLM 的 KL 惩罚项的梯度 eq. (48) 形式并不相同，直接将 diffusion 的 $\log\pi_\theta$ 和 $\log\pi_{ref}$ 带入 eq. (5) 只能得到 eq. (33)，并不能得到正确的 eq. (29)。

也可以举出反例，特别考察 $\mathcal{L}_{DDPO_{wrong}}$ 中的 KL 惩罚项：

$$
\begin{aligned}
&-\mathcal{L}_{DDPO_{wrong_{KL}}} \\
&=\frac{\beta}{4\sigma_t^4}\mathbb{E}_{\boldsymbol{c}\sim D,t\sim U(0,T)}\Big[\left(\|\boldsymbol{x}_{t-1}-\boldsymbol{\mu}_{ref}\left(\boldsymbol{x}_t,\boldsymbol{c},t\right)\|^2-\|\boldsymbol{x}_{t-1}-\boldsymbol{\mu}_{\theta_{old}}\left(\boldsymbol{x}_t,\boldsymbol{c},t\right)\|^2\right) \\
&\qquad\qquad\cdot\|\boldsymbol{x}_{t-1}-\boldsymbol{\mu}_{\theta}\left(\boldsymbol{x}_t,\boldsymbol{c},t\right)\|^2\Big]
\end{aligned}
\tag{32}
$$

其梯度表达式为：

$$
\begin{aligned}
&-\nabla_{\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,\boldsymbol{c},t)}\mathcal{L}_{DDPO_{wrong_{KL}}} \\
&=\frac{\beta}{4\sigma_t^4}\mathbb{E}_{\boldsymbol{c}\sim D,t\sim U(0,T)}\Big[\left(\|\boldsymbol{x}_{t-1}-\boldsymbol{\mu}_{ref}\left(\boldsymbol{x}_t,\boldsymbol{c},t\right)\|^2-\|\boldsymbol{x}_{t-1}-\boldsymbol{\mu}_{\theta_{old}}\left(\boldsymbol{x}_t,\boldsymbol{c},t\right)\|^2\right) \\
&\qquad\qquad\cdot\nabla_{\boldsymbol{\mu}_\theta}\|\boldsymbol{x}_{t-1}-\boldsymbol{\mu}_{\theta}\|^2\Big] \\
&=\frac{\beta}{4\sigma_t^4}\mathbb{E}\Big[SG\left(\|\boldsymbol{x}_{t-1}-\boldsymbol{\mu}_{ref}\|^2-\|\boldsymbol{x}_{t-1}-\boldsymbol{\mu}_{\theta}\|^2\right)\left(\boldsymbol{\mu}_\theta-\boldsymbol{x}_{t-1}\right)\Big]
\end{aligned}
\tag{33}
$$

构造反例：当 $\boldsymbol{\mu}_{ref}>\boldsymbol{x}_{t-1}>\boldsymbol{\mu}_\theta$ 且 $\|\boldsymbol{x}_{t-1}-\boldsymbol{\mu}_{ref}\|^2>\|\boldsymbol{x}_{t-1}-\boldsymbol{\mu}_\theta\|^2$ 时，可得 $-\nabla_{\boldsymbol{\mu}_\theta}\mathcal{L}_{DDPO_{wrong_{KL}}}<0$。这表明即使 $\boldsymbol{\mu}_{ref}>\boldsymbol{\mu}_\theta$，$\boldsymbol{\mu}_\theta$ 仍会朝减小的方向更新，存在明显逻辑悖论。

LLM DPO 的理论基础 eq. (77) 依赖于 eq. (7) 的有效性，而 eq. (7) 的有效性基于 eq. (5)，只有 eq. (5) 成立，才能推导出 LLM DPO 的最终公式 eq. (84)。因为扩散模型的 RLHF 场景中直接带入 $\log\pi_\theta$ 和 $\log\pi_{ref}$ 的 eq. (5) 不成立，且能举出反例，则直接带入 diffusion 的 $\log\pi_\theta$ 和 $\log\pi_{ref}$ 到 eq. (84) 得到的扩散 DPO 公式 eq. (34)（引自Diffusion-DPO）将不成立：

$$
\begin{aligned}
L(\theta)=&-\mathbb{E}_{(\boldsymbol{x}_0^w,\boldsymbol{x}_0^l)\sim\mathcal{D},t\sim\mathcal{U}(0,T),\boldsymbol{x}_t^w\sim q(\boldsymbol{x}_t^w|\boldsymbol{x}_0^w),\boldsymbol{x}_t^l\sim q(\boldsymbol{x}_t^l|\boldsymbol{x}_0^l)} \\
&\log\sigma\Bigg(-\beta T\omega(\lambda_t)\Bigg(\|\boldsymbol{\epsilon}^w-\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^w,t)\|_2^2-\|\boldsymbol{\epsilon}^w-\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^w,t)\|_2^2 \\
&-\left[\|\boldsymbol{\epsilon}^l-\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t^l,t)\|_2^2-\|\boldsymbol{\epsilon}^l-\boldsymbol{\epsilon}_{\text{ref}}(\boldsymbol{x}_t^l,t)\|_2^2\right]\Bigg)\Bigg)
\end{aligned}
\tag{34}
$$

3.4 修正的 DiffusionDPO

原始 DPO 梯度：

$$\nabla_\theta \mathcal{L}_{\mathrm{DPO}}(\pi_\theta; \pi_{\mathrm{ref}}) =$$

$$-\beta \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\underbrace{\sigma(\hat{r}_\theta(x, y_l) - \hat{r}_\theta(x, y_w))}_{\text{higher weight when reward estimate is wrong}} \left[\underbrace{\nabla_\theta \log \pi(y_w \mid x)}_{\text{increase likelihood of } y_w} - \underbrace{\nabla_\theta \log \pi(y_l \mid x)}_{\text{decrease likelihood of } y_l} \right] \right], \tag{35}$$

修正的 DiffusionDPO loss：

$$\mathcal{L}_{\mathrm{Diffusion_{DPO}}}(\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, \boldsymbol{c}, t); \boldsymbol{\mu}_{ref}(\boldsymbol{x}_t, \boldsymbol{c}, t)) =$$

$$-\beta \mathbb{E}_{(\boldsymbol{c},\boldsymbol{x}_t^w,\boldsymbol{x}_t^l)\sim\mathcal{D}} \left[\underbrace{\sigma(\hat{r}_{\theta_{old}}(\boldsymbol{c}, \boldsymbol{x}_t^l) - \hat{r}_{\theta_{old}}(\boldsymbol{c}, \boldsymbol{x}_t^w))}_{\text{higher weight when reward estimate is wrong}} \left[\underbrace{-\frac{\|\boldsymbol{x}_{t-1}^w - \boldsymbol{\mu}_\theta(\boldsymbol{x}_t^w, \boldsymbol{c}, t)\|^2}{2\sigma_t^2}}_{\text{increase likelihood of } \boldsymbol{x}_t^w} \right. \right.$$

$$\left. \left. + \underbrace{\frac{\|\boldsymbol{x}_{t-1}^l - \boldsymbol{\mu}_\theta(\boldsymbol{x}_t^l, \boldsymbol{c}, t)\|^2}{2\sigma_t^2}}_{\text{decrease likelihood of } \boldsymbol{x}_t^l} \right] \right], \tag{36}$$

$$\hat{r}_{\theta_{old}}(\boldsymbol{c}, \boldsymbol{x}_t) = \beta\left(\frac{\|\boldsymbol{x}_{t-1} - \boldsymbol{\mu}_{\mathrm{ref}}(x_t, c, t)\|^2}{2\sigma_t^2} - \frac{\|\boldsymbol{\mu}_{\theta_{old}}(x_t, c, t) - \boldsymbol{\mu}_{\mathrm{ref}}(x_t, c, t)\|^2}{2\sigma_t^2}\right)$$

A 分类模型 KL 散度梯度推导

A.1 KL 散度的定义

KL 散度衡量两个离散概率分布 π_θ 和 π_{ref} 之间的差异，定义为：

$$\mathrm{KL}(\pi_\theta(y|x) \| \pi_{ref}(y|x)) = \mathbb{E}_{x\sim D, y\sim\pi_\theta(y|x)} \left[\log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} \right] = \mathbb{E}_{x\sim D} \sum_{y\in\mathcal{Y}} \pi_\theta(y|x) \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}, \tag{37}$$

其中 $\mathcal{Y}$ 为离散类别空间。

A.2 KL 散度梯度推导步骤

步骤 1：展开 KL 表达式 直接写出离散求和形式：

$$\mathrm{KL}\big(\pi_\theta(\cdot|x) \| \pi_{ref}(\cdot|x)\big) = \sum_y \pi_\theta(y|x) \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}. \tag{38}$$

步骤 2：应用梯度算子 对 θ 求梯度：

$$-\nabla_\theta \mathrm{KL}\big(\pi_\theta(\cdot|x)\|\pi_{ref}(\cdot|x)\big) = -\nabla_\theta \sum_y \pi_\theta(y|x)\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}. \tag{39}$$

步骤 3：交换求和与梯度 由于求和项有限且 $\pi_\theta(y|x)$ 光滑，可交换求和与梯度：

$$-\nabla_\theta \mathrm{KL}\big(\pi_\theta(\cdot|x)\|\pi_{ref}(\cdot|x)\big) = -\sum_y \nabla_\theta\left[\pi_\theta(y|x)\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}\right]. \tag{40}$$

步骤 4：乘积法则分解 对每一项应用乘积法则：

$$-\nabla_\theta\left[\pi_\theta(y|x)\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}\right] = -[\underbrace{\nabla_\theta\pi_\theta(y|x)\cdot\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}}_{\text{Term 1}} + \underbrace{\pi_\theta(y|x)\cdot\nabla_\theta\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}}_{\text{Term 2}}]. \tag{41}$$

步骤 5：简化 Term 2 由于 $\pi_{ref}(y|x)$ 与 θ 无关，有：

$$\nabla_\theta\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} = \nabla_\theta\log\pi_\theta(y|x) = \frac{\nabla_\theta\pi_\theta(y|x)}{\pi_\theta(y|x)}. \tag{42}$$

因此 Term 2 简化为：

$$\pi_\theta(y|x)\cdot\frac{\nabla_\theta\pi_\theta(y|x)}{\pi_\theta(y|x)} = \nabla_\theta\pi_\theta(y|x). \tag{43}$$

步骤 6：合并两项 将 Term 1 和 Term 2 相加：

$$\sum_y\left[\nabla_\theta\pi_\theta(y|x)\cdot\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)} + \nabla_\theta\pi_\theta(y|x)\right]. \tag{44}$$

步骤 7：处理归一化条件 由于 $\sum_y \pi_\theta(y|x) = 1$，其梯度为 0：

$$\sum_y \nabla_\theta\pi_\theta(y|x) = \nabla_\theta\sum_y\pi_\theta(y|x) = \nabla_\theta 1 = 0. \tag{45}$$

因此第二项求和为 0，仅保留第一项：

$$-\nabla_\theta \mathrm{KL}\big(\pi_\theta(\cdot|x)\|\pi_{ref}(\cdot|x)\big) = -\sum_y \nabla_\theta\pi_\theta(y|x)\cdot\log\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}. \tag{46}$$

步骤 8：对数导数技巧 利用 $\nabla_\theta \pi_\theta(y|x) = \pi_\theta(y|x)\nabla_\theta \log \pi_\theta(y|x)$，改写为：

$$-\nabla_\theta \mathrm{KL}\big(\pi_\theta(\cdot|x)\|\pi_{ref}(\cdot|x)\big) = -\sum_y \pi_\theta(y|x)\nabla_\theta \log \pi_\theta(y|x) \cdot \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}. \tag{47}$$

步骤 9：期望形式 最终梯度可表示为期望：

$$-\nabla_\theta \mathrm{KL}(\pi_\theta(\cdot|x)\|\pi_{ref}(\cdot|x)) = -\mathbb{E}_{x\sim D, y\sim\pi_\theta(y|x)}\left[\nabla_\theta \log \pi_\theta(y|x) \cdot \log \frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}\right]. \tag{48}$$

k1 损失函数形式 根据梯度公式 eq. (48)，损失函数可推导为：

$$\mathcal{L}_{\mathrm{KL}_t}(\pi_\theta, \pi_{\mathrm{ref}}) = \mathbb{E}_{x\sim D, y\sim\pi_{\theta_{old}}(y|x)}\left[\log \frac{\pi_{\theta_{old}}(y|x)}{\pi_{\mathrm{ref}}(y|x)} \cdot \log \pi_\theta(y|x)\right] \tag{49}$$

k2 损失函数形式 KL 梯度

$$-\nabla_\theta \mathrm{KL}(\pi_\theta(\cdot|x)\|\pi_{ref}(\cdot|x)) = g_{\mathrm{KL}}(\theta) \tag{50}$$

$$= -\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[\nabla_\theta \log \pi_\theta(y|x) \log \frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right] \tag{51}$$

$$-\nabla_\theta \mathrm{KL}(\pi_{\theta_{old}}(\cdot|x)\|\pi_{ref}(\cdot|x)) = g_{\mathrm{KL}}(\theta_{\mathrm{old}}) \tag{52}$$

$$= -\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_{\mathrm{old}}(\cdot|x)}\left[\nabla_\theta \log \pi_{\mathrm{old}}(y|x) \cdot \log \frac{\pi_{\mathrm{old}}(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right] \tag{53}$$

$$= -\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_{\mathrm{old}}(\cdot|x)}\left[\nabla_\theta \log \pi_\theta(y|x)|_{\theta=\theta_{\mathrm{old}}} \cdot \log \frac{\pi_{\mathrm{old}}(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right] \tag{54}$$

$$= -\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_{\mathrm{old}}(\cdot|x)}\left[\nabla_\theta \log \frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}\Big|_{\theta=\theta_{\mathrm{old}}} \cdot \log \frac{\pi_{\mathrm{old}}(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right] \tag{55}$$

$$= -\frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_{\mathrm{old}}(\cdot|x)}\left[\nabla_\theta \left(\log \frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right)^2\Big|_{\theta=\theta_{\mathrm{old}}}\right] \tag{56}$$

定义 k_2 loss 下的损失函数

$$\mathcal{L}_{k_2}(\theta) := -\frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_{\mathrm{old}}(\cdot|x)}\left[\left(\log \frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right)^2\right] \tag{57}$$

求损失函数的梯度

$$\nabla_\theta \mathcal{L}_{k_2}(\theta) = -\frac{1}{2}\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_{\text{old}}(\cdot|x)}\left[\nabla_\theta\left(\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right)^2\right] \tag{58}$$

$$= -\mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_{\text{old}}(\cdot|x)}\left[\nabla_\theta\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\cdot\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right] \tag{59}$$

得出结论

$$\nabla_\theta \mathcal{L}_{k_2}(\theta_{\text{old}}) = g_{\text{KL}}(\theta_{\text{old}}) \tag{60}$$

B 高维高斯分布的 KL 散度和梯度的推导

考虑两个 k 维各向同性高斯分布：$P\sim\mathcal{N}(\boldsymbol{\mu}_\theta,\sigma_1^2\mathbf{I})$, $Q\sim\mathcal{N}(\boldsymbol{\mu}_{\text{ref}},\sigma_2^2\mathbf{I})$
其概率密度函数分别为：

$$\begin{aligned} P(\mathbf{x}) &= (2\pi\sigma_1^2)^{-k/2}\exp\left(-\frac{\|\mathbf{x}-\boldsymbol{\mu}_\theta\|^2}{2\sigma_1^2}\right)\\ Q(\mathbf{x}) &= (2\pi\sigma_2^2)^{-k/2}\exp\left(-\frac{\|\mathbf{x}-\boldsymbol{\mu}_{\text{ref}}\|^2}{2\sigma_2^2}\right)\end{aligned} \tag{61}$$

B.1 KL 散度定义

KL 散度定义为：

$$D_{\text{KL}}(P\parallel Q) = \mathbb{E}_{\mathbf{x}\sim P}\left[\ln\frac{P(\mathbf{x})}{Q(\mathbf{x})}\right] \tag{62}$$

将密度函数代入，展开对数比：

$$\ln\frac{P(\mathbf{x})}{Q(\mathbf{x})} = \frac{k}{2}\ln\frac{\sigma_2^2}{\sigma_1^2} + \left[-\frac{\|\mathbf{x}-\boldsymbol{\mu}_\theta\|^2}{2\sigma_1^2} + \frac{\|\mathbf{x}-\boldsymbol{\mu}_{\text{ref}}\|^2}{2\sigma_2^2}\right] \tag{63}$$

B.2 逐项计算期望

将 KL 散度拆分为常数项与二次项之和：

$$D_{\text{KL}} = \underbrace{\frac{k}{2}\ln\frac{\sigma_2^2}{\sigma_1^2}}_{\text{常数项}} + \underbrace{\mathbb{E}_P\left[-\frac{\|\mathbf{x}-\boldsymbol{\mu}_\theta\|^2}{2\sigma_1^2} + \frac{\|\mathbf{x}-\boldsymbol{\mu}_{\text{ref}}\|^2}{2\sigma_2^2}\right]}_{\text{二次项}} \tag{64}$$

计算 $\mathbb{E}_P[\|\mathbf{x}-\boldsymbol{\mu}_\theta\|^2]$　由于 $\mathbf{x}\sim P$，协方差矩阵为 $\sigma_1^2\mathbf{I}$，故：

$$\mathbb{E}_P[\|\mathbf{x}-\boldsymbol{\mu}_\theta\|^2] = \text{tr}(\sigma_1^2\mathbf{I}) = k\sigma_1^2 \tag{65}$$

计算 $\mathbb{E}_P[\|\mathbf{x}-\boldsymbol{\mu}_{\mathbf{ref}}\|^2]$ 展开二次型并取期望：

$$\begin{aligned}\|\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ref}}\|^2 &= \|\mathbf{x}-\boldsymbol{\mu}_\theta+\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2 \\ &= \|\mathbf{x}-\boldsymbol{\mu}_\theta\|^2+2(\mathbf{x}-\boldsymbol{\mu}_\theta)^\top(\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}})+\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2\end{aligned} \tag{66}$$

取期望后，交叉项因 $\mathbb{E}_P[\mathbf{x}-\boldsymbol{\mu}_\theta]=0$ 而消失：

$$\mathbb{E}_P[\|\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ref}}\|^2]=k\sigma_1^2+\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2 \tag{67}$$

代入二次项

$$\begin{aligned}\mathbb{E}_P\left[-\frac{\|\mathbf{x}-\boldsymbol{\mu}_\theta\|^2}{2\sigma_1^2}+\frac{\|\mathbf{x}-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2}\right] &= -\frac{k\sigma_1^2}{2\sigma_1^2}+\frac{k\sigma_1^2+\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2} \\ &= -\frac{k}{2}+\frac{k\sigma_1^2}{2\sigma_2^2}+\frac{\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2}\end{aligned} \tag{68}$$

合并所有项 将常数项与二次项合并：

$$\begin{aligned}D_{\mathrm{KL}} &= \frac{k}{2}\ln\frac{\sigma_2^2}{\sigma_1^2}-\frac{k}{2}+\frac{k\sigma_1^2}{2\sigma_2^2}+\frac{\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2} \\ &= \frac{k}{2}\left[2\ln\frac{\sigma_2}{\sigma_1}+\frac{\sigma_1^2}{\sigma_2^2}-1\right]+\frac{\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2}\end{aligned} \tag{69}$$

最终公式 整理后得到：

$$D_{\mathrm{KL}}(P\parallel Q)=\underbrace{k\left(\ln\frac{\sigma_2}{\sigma_1}+\frac{\sigma_1^2}{2\sigma_2^2}-\frac{1}{2}\right)}_{\text{方差项}}+\underbrace{\frac{\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2}}_{\text{均值项}} \tag{70}$$

当 $\sigma=\sigma_1=\sigma_2$ 时，

$$D_{\mathrm{KL}}(P\parallel Q)=\frac{\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma^2} \tag{71}$$

B.3 对 $\boldsymbol{\mu}_\theta$ 的梯度推导

从 KL 散度公式中提取与 $\boldsymbol{\mu}_\theta$ 相关的项：

$$D_{\mathrm{KL}}(P\parallel Q)=\underbrace{\frac{\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_{\mathrm{ref}}\|^2}{2\sigma_2^2}}_{\text{唯一与 }\boldsymbol{\mu}_\theta\text{ 相关的项}}+\text{其他与 }\boldsymbol{\mu}_\theta\text{ 无关的项.} \tag{72}$$

步骤 1：展开二次型 对 $\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|^2$ 展开：

$$\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}\|^2 = (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}})^\top (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}). \tag{73}$$

步骤 2：计算关于 $\boldsymbol{\mu}_\theta$ 的梯度 利用二次型的梯度公式：

$$-\nabla_{\boldsymbol{\mu}_\theta} \left[(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}})^\top (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}) \right] = -2(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}). \tag{74}$$

步骤 3：组合梯度分量 将梯度结果代入 KL 散度表达式：

$$-\nabla_{\boldsymbol{\mu}_\theta} D_{\text{KL}} = -\frac{1}{2\sigma_2^2} \cdot 2(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}) = -\frac{\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}}{\sigma_2^2}. \tag{75}$$

最终梯度表达式

$$-\nabla_{\boldsymbol{\mu}_\theta} D_{\text{KL}}(P \parallel Q) = -\frac{\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_{\text{ref}}}{\sigma_2^2} \tag{76}$$

C DPO 推导

本节基于 KL 约束的奖励最大化目标，推导出可操作的直接偏好优化目标函数。首先建立基础优化问题：

$$\begin{aligned}
&\max_\pi \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi}[r(x,y)] - \beta D_{\text{KL}}[\pi(y|x) \| \pi_{\text{ref}}(y|x)] \\
&= \max_\pi \mathbb{E}_{x\sim\mathcal{D}} \mathbb{E}_{y\sim\pi(y|x)} \left[r(x,y) - \beta \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} \right] \\
&= \min_\pi \mathbb{E}_{x\sim\mathcal{D}} \mathbb{E}_{y\sim\pi(y|x)} \left[\log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} - \frac{1}{\beta} r(x,y) \right]
\end{aligned} \tag{77}$$

为求解该优化问题，引入配分函数 $Z(x)$ 构造概率分布：

$$Z(x) = \sum_y \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x,y) \right) \tag{78}$$

定义新的参考分布 π^* 为：

$$\pi^*(y|x) = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp\left(\frac{1}{\beta} r(x,y) \right) \tag{79}$$

将目标函数重构为：

$$
\begin{aligned}
&\min_{\pi} \mathbb{E}_{x\sim\mathcal{D},y\sim\pi(y|x)} \left[\log \frac{\pi(y|x)}{\pi^*(y|x)} - \log Z(x)\right] \\
&= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}} \left[\mathbb{E}_{y\sim\pi(y|x)} \left[\log \frac{\pi(y|x)}{\pi^*(y|x)}\right] - \log Z(x)\right] \\
&= \min_{\pi} \mathbb{E}_{x\sim\mathcal{D}} \left[D_{KL}\left(\pi(y|x) \parallel \pi^*(y|x)\right) - \log Z(x)\right]
\end{aligned} \tag{80}
$$

由于 $Z(x)$ 与策略 π 无关，优化目标简化为最小化 KL 散度项。根据 KL 散度的非负性，当 $\pi = \pi^*$ 时取得全局最优解。

C.1 从奖励建模到偏好学习

实际应用中直接求解 π^* 存在两大障碍：1) 真实奖励函数 r^* 未知；2) 配分函数 $Z(x)$ 的计算涉及全响应空间积分。为此，我们引入偏好学习框架。

采用 Bradley-Terry 模型，对于输入 x 和响应对 (y_w, y_l)，偏好概率建模为：

$$
p^*(y_w \succ y_l|x) = \frac{\exp(r^*(x, y_w))}{\exp(r^*(x, y_w)) + \exp(r^*(x, y_l))} \tag{81}
$$

关键突破在于建立奖励函数与最优策略的显式关联。由式 (2) 可得：

$$
r(x, y) = \beta \log \frac{\pi^*(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x) \tag{82}
$$

将奖励差表达式代入偏好概率模型：

$$
p^*(y_w \succ y_l|x) = \frac{1}{1 + \exp\left(\beta \log \frac{\pi^*(y_l|x)}{\pi_{\text{ref}}(y_l|x)} - \beta \log \frac{\pi^*(y_w|x)}{\pi_{\text{ref}}(y_w|x)}\right)} \tag{83}
$$

C.2 最终目标函数

通过极大似然估计，得到直接优化策略的参数化目标函数：

$$
\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \tag{84}
$$

D 重要性采样

D.1 基本概念与动机

重要性采样（Importance Sampling）是强化学习中实现离策略（off-policy）学习的关键技术，其核心思想是通过引入行为策略（behavior policy）的采样分布来估计目标策略（target policy）的期望值。这一方法在策略优化中具有双重意义：

1. ** 样本复用 **：允许利用历史策略生成的旧样本进行当前策略更新，显著提升数据利用率 2. ** 方差控制 **：通过重要性权重修正新旧策略的概率分布偏差，维持无偏估计特性

D.2 策略梯度推导

考虑策略梯度基本形式：

$$\nabla_\theta J(\theta) = \mathbb{E}_{(s_t,a_t)\sim\pi_\theta}\left[A_\theta(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)\right] \tag{85}$$

当转换为离策略更新时，需要引入重要性权重（Importance Weight）$\rho_t = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta'}(a_t|s_t)}$：

$$\begin{aligned}
\nabla_\theta J(\theta) &= \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{\pi_\theta(s_t,a_t)}{\pi_{\theta'}(s_t,a_t)}A_\theta(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)\right] \\
&= \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{\pi_\theta(a_t|s_t)\pi_\theta(s_t)}{\pi_{\theta'}(a_t|s_t)\pi_{\theta'}(s_t)}A_\theta(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)\right] \\
&\approx \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta'}}\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta'}(a_t|s_t)}A_{\theta'}(s_t,a_t)\nabla_\theta \log \pi_\theta(a_t|s_t)\right]
\end{aligned} \tag{86}$$

推导过程中包含两个重要近似：

- 状态分布抵消假设：$\pi_\theta(s_t) \approx \pi_{\theta'}(s_t)$，在策略更新幅度较小时成立
- 优势函数近似：$A_\theta(s_t,a_t) \approx A_{\theta'}(s_t,a_t)$，要求新旧策略差异可控

D.3 目标函数形式化

令 $\theta' = \theta_{\text{old}}$，得到离策略目标函数：

$$J(\theta) = \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta_{\text{old}}}} \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} A_{\theta_{\text{old}}}(s_t, a_t) \right] \tag{87}$$

该目标函数具有以下特性：

1. **无偏性**：当 π_θ 与 $\pi_{\theta_{\text{old}}}$ 的支撑集相同时保持无偏估计
2. **方差敏感性**：重要性权重 ρ_t 的数值稳定性直接影响梯度质量
3. **策略约束**：需通过 KL 散度等度量限制 π_θ 与 $\pi_{\theta_{\text{old}}}$ 的差异